# Application of Visual Analytic and Algorithmic Techniques for dealing with Multidimensional Datasets: A Literature Review

Daniel Dixey, MSc Data Science, City University London

**Abstract—**The vast amount of data and computational power readily available now means that we are now in a period of data explosion. This has been caused by the declining cost of components and the vast number of new technologies available to make storage and processing easier and faster. With all this computational power available to us as humans, understanding multidimensional datasets has become a significantly more central task in analyses and predictive modelling. This literature review critically assesses three methods of dealing with this type of data, via a systematic framework, in conjunction with various dimension reduction algorithms. It is concluded that all papers present viable solutions for dealing with multidimensional datasets through the cooperation of computational techniques and human interpretation.

**Index Terms**— Literature Review, Visual analytics, Multidimensional data, VHDR, Autoencoder

## INTRODUCTION

There is a growing need to extract understanding and predictive value from datasets. In order to facilitate such tasks, the use of visual analytics has become an important step in both exploratory analysis and presentation of results to stakeholders. With the proliferation of number of packages and tools increasing year on year, the price of hardware and the rise in the number of devices collecting data is in a period of exponential growth. The need to obtain value by selecting the correct visual is important but more essentially, reducing the number of dimensions has become an important process step too.

This literature review represents a comprehensive analysis of three papers which all discuss dimension reduction techniques for dealing with multidimensional datasets. The paper is organised as follows, firstly; the challenges of dealing with multidimensional data in relation to the computational and human difficulties are discussed, secondly; a synopsis of two case studies from the literature are presented and finally a concise discussion, a conclusion and any implications of implementing the different techniques discussed are presented.

The papers that form the basis of this literature review cover three methods of dealing with multidimensional datasets; Visual Hierarchical Dimension Reduction (VHDR) [3], Kuntal et al.'s use of the Igloo-Plot framework and lastly "Reducing the Dimensionality of Data with Neural Networks" by Hinton et al [2].

The topic of multidimensional data covers a comprehensive number of fields and domains. A multidimensional dataset is such that the number of dimensions included is greater than two. Human intelligence means a table of data up to three dimensions can usually be dealt with without too much difficulty in a visual context; we can interpret surfaces and objects in their principal dimensions. At four dimensions the visual begins to lose its value and anything with greater than five dimensions is almost impossible for humans to visualise. It is the visual cortex in the brain which limits human ability to see no more than 3 dimensions. It is concluded by Rehmeyer [7] that it is the distance between the eyes of both which causes human and animal to only see images in two-dimensions. It is only with the combination of two eyes and a visual cortex that a human or animal is able to interpret the space in front of them in three dimensions. This limitation means the careful and challenging selection of only 2-3 features to represent an entire data set.

When you consider that modern datasets have many hundreds, if not thousands, of features, there is no feasible means of representing the relationship of each of the features simultaneously on a visual plot which is interpretable by simple human intelligence.

There are many visualisation techniques that have tried to overcome the implications of dealing with such large number of dimensions; these include glyph techniques, parallel coordinates and scatter plots [8]. These methods, as well developed as they are, lack the complexity to be able to create informative visuals when refined for the large datasets, therefore it is vital to consider both the algorithmic and human challenges of dealing datasets of this magnitude.

A further limitation to address is the selection of an appropriate approach for choosing a dimension reducing algorithm. There are three well used methods for dimension reductions; Principal Component Analysis (PCA) [4], Multidimensional Scaling (MDS) [5] and Self Organising Maps (SOMs) [6]. Computationally all the methods will work will most datasets, but unfortunately, each of these methods will provide, in most cases, a slightly biased set of results.

There are three main considerations to acknowledge when applying feature selection. Firstly, applying feature selection reduction too early on in the exploratory analysis phase could be detrimental to the overall result or data product. This is because the most interesting patterns for analysis and prediction may be hidden in deep subspaces not accessible if selection is made prematurely. This limitation also inhibits the user whereby the choice of variables can limit the ability to discover patterns outside of the users expertise. The second consideration is that the reduction algorithms are not unique or specific for one particular domain and therefore in some cases the interpretation can be quite abstract for the analyst attempting to extract meaning. The third consideration, which will not be considered in this literature review, is the analysis of the various algorithms capable of dimension reduction as this will either inhibit or exhibit the underlying patterns in lower sub spaces of multidimensional data.

## CASE STUDIES

This review will focus on the case studies from two papers; "Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets" by Yang et al [3] and "Igloo-Plot: A tool for visualization of multidimensional datasets" by Kuntal et al. [2].

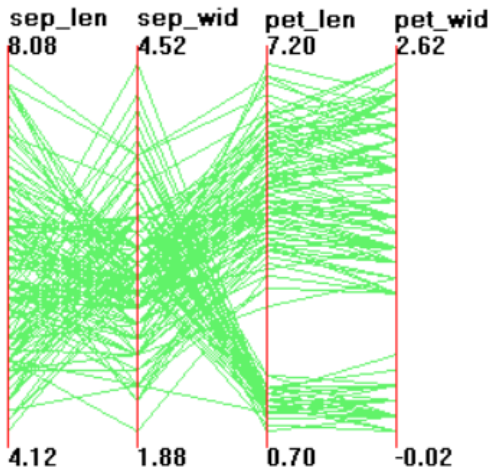


Fig.1. Iris Data set represented in a Parallel coordinates plot (4 dimensions shown).

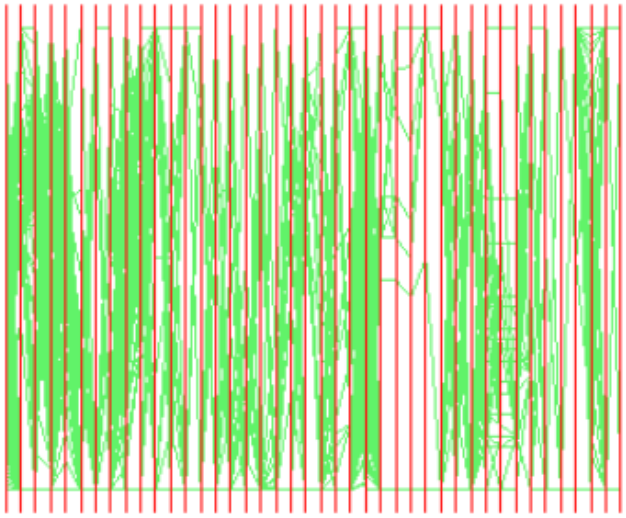

Fig.2. Census income data represented in a Parallel coordinates plot (43 dimensions shown).

## Case Study: Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets

Presented in the introduction of the paper [3] the dataset, census data, and the problem of multidimensional data visualisation is discussed. Yang et al., describe the immediate difficulty of working with multiple dimensional datasets succinctly by comparing to the well know Iris dataset. Through a comparison of two different data sets, a visualisation of a parallel coordinates plot is made use of. "Figure 2 shows a subset of the Census income data set, which ***h***as 42 dimensions and 200 data items. While the number of data items in this display is comparable to Figure 1, individual data items can no longer be seen clearly for this display, since the number of dimensions has been greatly increased" [10]. The use of parallel coordinates is used well throughout the paper to present the difficult whereby the underlying intention is used well as a means of comparison and also to show the reader how the framework can be used.

In Section 8 of the paper, the application of the VHDR framework has been described when applied to PUMS Census data. The data has been collected from the Los Angeles and Long Beach areas in 1970, 1989 and 1990 [9], where the data consists of 42 dimensions each with 200 instances. No data transformations have been described in the paper and since the dataset was utilised in another study (IPUMS project) [9], that all typical pre-processing and normalisation techniques been implemented in the dataset in order to get it into an aggregated state.

The framework has been developed to form the tool "Xmdvtool", the work of Prof. Matthew Ward. The tool supports the well established technique and important visualisation technique called "Brushing" to aid knowledge discovery. As summarised by Martin et al. [10], "Brushing is an operation found in many data visualization systems. It is a mechanism for interactively selecting subsets of the data so that they may be highlighted, deleted, or masked." In addition to this, the tool utilises proximity based colouring and a variety of other filtering and visualisation tools such as dimension zooming, reordering and drill-down capability which provides the user the ability to assist with better understanding of the data. The functionality within the tool has been derived from the framework, described in the case study the steps taken have been described and the reasoning why. While there is more emphasis on how the visualisations support human reasoning, which is described in depth throughout the paper, it is not depicted in relation to the dataset.

As shown in figure 6, below**,** the dataset has been projected onto a lower dimensional space. When identifying different clusters the authors decided to supplement the computational clustering methods with domain knowledge, this is expressed in the final paragraph of Section 8. The progressive journey is visualised well in the final figure of the paper figure 6**,** it can be seen quite clearly how the division of work has been supplemented by the knowledge of the authors after the use of an automatic clustering technique (from left to right of the figure).

## Case Study: Igloo-Plot: A tool for visualization of multidimensional datasets

The paper [1] describes a systematic approach to visualising multidimensional datasets using an Igloo Plot. The Igloo Plot can be thought of as the projection of data-points on to a semi-circle plot. The purpose of projecting the points onto this two-dimensional space is that it is possible to visualize the spatial correlations of the variables when clustering, string connections for determining feature importance and also a variation plot.

The analytical strategy that has been created can be broken down into six distinct phases. At each phase the work is divided equally between the human and the algorithms within the dedicated 'Igloo-Plot' tool. The first stage is where the user interacts with the data in the input file/table. The second stage is where the user interacts with the data through a series of basic visuals to access the data-points; the tool does not recommend normalization however it is important to note that if you would like to apply a z-score transformation, normalization is then suggested. Steps 3 to 5 calculate the correlations sorted by rank (highly correlated features are grouped together) and finally the data is projected onto a 2 dimensional semi-circle space. The final step in the process is one that plots the projection into a worthwhile visualization.

Computational algorithms ~~that~~ are employed by the tool; z-score for the normalization of features and the pair-wise correlation coefficient calculates to determine the relationship between each of the features and finally in order to project the correlation matrix on to a 2 dimensional space utilizes Hooke's law. The output of the pair-wise calculation is a square correlation matrix, Euclidean or correlation, and the tool has the inbuilt functionality to transform that matrix onto visual sub space.

In order to extract value from the visualization, the interactive functionality allows the end-user to select and highlight specific groups or points within the visualization. Further to this, the tool also supports the ability to remove features, this can be completed in any of the three types of 'Igloo' plots; data-points, string connections and variation plot, or heat

Fig. 3. Feature reduction by Latent Semantic Analysis example (LSA).

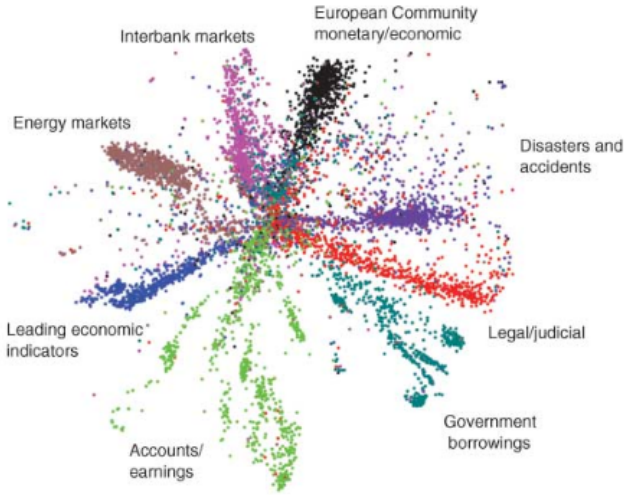


Fig. 4. The same dataset as shown in Figure 3 however this scatter plot is generated by a Autoencoder Neural Network [2].

maps. The intention and benefit of having a variety of plots is such that the user of the tool is able to select the most appropriate to impart understanding and generate insights about the complexity of the data of each different data set.

Throughout the paper three unrelated datasets, Olives, Wines and the Gene-Expression dataset, are used to demonstrate the potential of the Igloo Plot. Each dataset is used in turn to show how the user derives insights and how each variety of plots supports a particular method. Figure 8 from the paper (shown below) demonstrates, with the aid of a colouring technique, protein consumption by country via a string connection igloo plot. The strings represents, according to the colour level, the similarity between each of the types of food. The use of spatial relationship is also supported by colour which complements the distance relationships. The colouring aids human interpretation, such that it is easy to identify strong and weak relationships, with the addition of filtering for a specified range.

## DISCUSSION

VHDR [3] and the Igloo [2] methodologies have attempted to differentiate themselves from traditional methods of dimension reduction. This has been made possible through the exploitation of the available interaction between the human user and the different DR algorithms. The framework of these methods has been in existence since 1994 [12] and many papers have been written that utilised the framework extensively for the analysis.

The use of a hierarchical dimension cluster tree shows how the analyst can interrogate the data and have his/her own influence on how the clustering algorithms behave at lower dimensional spaces. The methodology that is described is not rigid in terms of following a process; it is however described as one that is fluid, whereby the exploration can change easily between its various tasks. The advantages of such fluidity are that the methodology is generic and one that could be applied to a multitude of different disciplines.

One advanced method of dimension reduction that utilises Neural Networks is presented by Hinton et al. [2]. With the use of autoencoder networks it is possible to train a neural network to recognise and learn lower dimensional spaces. The advantages of using a neural network are; it is very good for non-linear dimension reduction. it is possible to decode the projection for further analysis/further reduction and finally the visualisations that are produced as a result can be potentially better for user interpretation.

The table below presents the most distinct points about each method for comparison between each method.

| | 1. VHDR | 2. Igloo-Plot | 3. Autoencoder Neural Networks (ANN) |
|---|---|---|---|
| **Uncertainty in Data / Limitations / Generalization** | A: This framework would scale well as it has been designed such that it is generic.<br><br>B: No discussion with how to deal with noisy data or outliers. | A: The tool's first step for analysing the dataset makes use of descriptive and simple plots to understanding the characteristics, noise and outlier detection could be detected here.<br><br>B: A limitation that is not expressed clearly is how the tool deals with categorical data.<br><br>C: Tool is not fixed to one specific domain. | A: This method of DR could be applied to a large variety of dataset provided that there is sufficient instances and labels.<br><br>B: This method will learn the underlying characteristics of the data. When training the network you are essentially training the network to map itself, therefore it should not learn the noise.<br><br>C: Method is applicable for most domains given that the data is transformed correctly. |
| **Complexity of Approaches** | A: Multiple transformations are necessary in succession to find the most appropriate method. This in turn then lengthens the scale of the VA task. | A: Ascertaining if normalisation is required at the initial stage. Structure of the input data, no discussion has mentioned the input of the data is required to be in a particular form. | A: Determining optimised coefficients is difficult when tuning the neural network.<br>B: Method scales well for Big Data - If you are dealing with large datasets and it means that its not computation viable to |
| **Visual Clutter / Layout** | A: A well developed framework that utilised three methods of dimension reduction to reduce the impact of visual clutter.<br><br>B: No discussion of how to deal with data that exhibits a non-linear behaviour. | A: Projection is mapped onto the semicircle plot. The Igloo-Plot tool utilises functionality that determines how to colour (VIBGYOR) and representation (String Connections, Variation Plots and Scatter plotting).<br><br>B: No discussion to advise if other traditional plotting is possible. | A: As visually shown in Figure 4B and 4C, the clutter of two different dimension reduction algorithms is shown. The NN work produces a less cluttered representation of the same dataset when compared to the Latent Semantic Analysis (LSA). |
| **Interpretation, validation and Final Points** | A: Heavy reliance of the knowledge and ability to of the user to identify the most useful subspaces for analysis through repeated comparisons of DR and VA. | A: Decision making. Interpretation of/interaction with the novel method visualisation.<br><br>B: Domain knowledge may be required to assist greatly with the interpretation of the variety of plots available. | A: As concluded in the paper the advantage of using a ANN is that it is possible to generate mapping of the data and code in which is a significant advantage over current non-parametric dimension reduction algorithms. |

## CONCLUSION

The papers which have been discussed in this literature review have focused on the challenges faced when using visualization and visual analytics of multidimensional data, specifically the inability to present much information, therefore the potential loss of information, when using dimension reduction. The different approaches, frameworks and visual and analytical tools have been discussed as part of this literature review.

The authors focus on applying a variety of different DR techniques to explore the data with the particular focus on uncovering interesting patterns in lower dimensional plots. A variety of different of interactive visual tools have been created and customised for application to each particular framework, specifically VHDR and Igloo-plots.

All of the approaches reviewed are generic by design and therefore could readily be applied to different domains or extended further by supplementary research. In summary, the papers offer

analogous solutions in dealing with multidimensional datasets with the aid of computation and human cooperation and all face similar limitations.

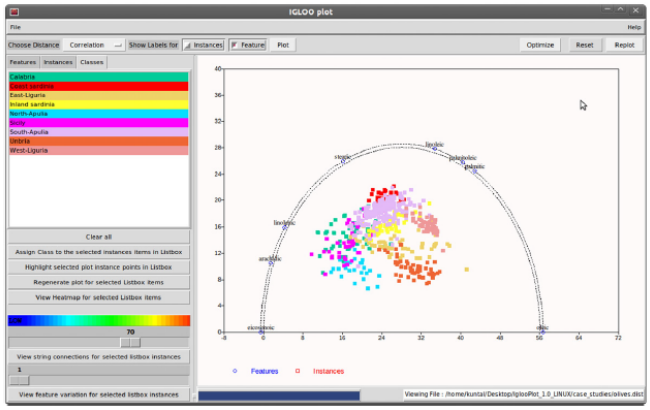


Fig. 5. A demonstration of the Igloo-Plot, the figure displays the Wine dataset. Features are shown on the edge of the plot (blue text), the distance of each instance to a features and other can be interpreted as the similarity between features and other instances.

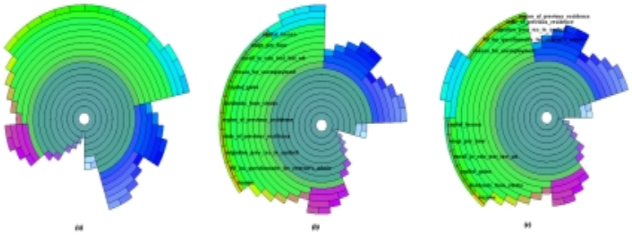

Fig. 6. Dimension hierarchy of the Census dataset in InterRing. Figure (a) shows the automatically generated hierarchy. Figure (b) shows the detail of a cluster after brushing and rotation. Figures (c) shows the modified hierarchy after moving some dimensions from that cluster to elsewhere.

## ACKNOWLEDGMENTS

The author wishes to thank the School of Mathematics, Computer Science & Engineering at City University for the opportunity to study towards a Masters in Data Science as a Part-Time Student.

## REFERENCES

[1] Bhusan K. Kuntal, Tarini Shankar Ghosh and Sharmila S. Mande, “Igloo-Plot: A tool for visualization of multidimensional datasets”, Genomics, Volume 103, Issue 1, Pages 11-20, January 2014.

[2] George E. Hinton and Rusian R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, Science, 313, 504-507, 2006.

[3] Jing Yang, Matthew O. Ward, Elke A. Rundensteiner and Shiping Huang, "Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets", VisSym, 2003.

[4] Ian Jolliffe, “Principal component analysis. “, Wiley Online Library, 2005.

[5] Warren S. Torgerson, "Multidimensional scaling: I. Theory and method.", Psychometrika,17, no. 4 (1952): 401-419, December 1952.

[6] Arthur Flexer, "On the use of self-organizing maps for clustering and visualization...” Intell. Data Anal.5, no. 5 (2001): 373-384, January 2001.

[7] Julie Rehmeyer, "Seeing in Four Dimensions." Science News. August 22, 2008. Accessed February 28, 2015. https://www.sciencenews.org/article/seeing-four-dimensions.

[8] Tze-Haw Huang, Mao L. Huang, and Kang Zhang, "An Interactive Scatter Plot Metrics Visualization for Decision Trend Analysis...” ICMLA (2), 2012.

[9] Steven Ruggles, "Sample Designs and Sampling Errors", Historical Methods. Volume 28. Number 1. Pages 40 – 46, 1995.

[10] Allen R. Martin and Matthew O. Ward, “High Dimensional Brushing for Interactive Exploration of Multivariate Data”, IEEE Computer Society, Washington USA, 1995.

[11] Matthew O. Ward, "Xmdvtool: Integrating Multiple Methods for Visualizing Multivariate Data," IEEE Conf. on Visualization '94, pp 326 - 333, Oct. 1994.

[12] Matthew Ward, "Xmdvtool Home Page: Documents." Xmdvtool Home Page: Documents, 2014. Accessed February 28, 2015. http://davis.wpi.edu/xmdv/documents.html.

[13] Alfred Inselberg and Bernard Dimsdale, “Parallel coordinates: a tool for visualizing multi-dimensional geometry”, Proceedings of the 1st conference on Visualization '90, October 23-26, San Francisco, California, 1990.